Tutoriat PS 6: Revision, new discrete Random Variables and transition to statistics

Mihai Duzi, Lucan Cristian, Luminaru Ionut

December 5, 2025

Contents

1 Variabile aleatoare specifice seriei 24

1.1 Distribuția Hipergeometrică

Distribuția hipergeometrică $X \sim \text{Hip}(N, M, n)$

- Se consideră problema extragerii repetate dintr-o cutie ce conține N obiecte, din care M sunt defecte.

- Dacă extragerile se fac **cu înlocuire** (obiectul extras este pus înapoi în cutie înainte de extragerea următoare), atunci numărul de obiecte defecte extrase în n extrageri este o variabilă aleatoare **binomială** cu parametrii n și $p = \frac{M}{N}$ (probabilitatea extragerii unui obiect defect).
- Dacă extragerile se fac **fără înlocuire**, atunci probabilitatea extragerii unui obiect defect nu mai este aceeași în cele n extrageri și deci în acest caz numărul de obiecte defecte extrase nu mai este o variabilă aleatoare binomială.

$\Rightarrow$ Se obține o variabilă aleatoare având funcția de probabilitate:

$$P(X=k) = \begin{cases} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, & \text{dacă } k \in \{\max(0, n-(N-M)), \dots, \min(n, M)\} \\ 0, & \text{în rest} \end{cases}$$

Aceasta se numește **distribuție hipergeometrică** cu parametrii N (populația totală), M (numărul de obiecte de tipul dorit în populație) și n (mărimea eșantionului).

Aplicarea Distribuției Hipergeometrice

Exercițiul 7. La un examen participă 100 de studenți, dintre care 5 copiază. Se realizează în sală 3 verificări simultane. Notăm cu X variabila aleatoare ce indică numărul de studenți depistați cu fraudă din cele trei verificări. Determinați:

a) Repartiția variabilei aleatoare X.

b) Probabilitatea ca toți cei trei studenți verificați să fi fost fraudulenți știind că cel puțin unul dintre ei a fost prins copiind.

Rezolvare. Fie N numărul total de studenți din care extragem ($N = 100$), N_1 este numărul studenților care copiază ($N_1 = 5$), N_2 este numărul studenților care nu copiază ($N_2 = 95$). Pentru o extragere de n studenți ($n = 3$), n_1 este numărul studenților găsiți că copiază, n_2 este numărul studenților care nu copiau, dintre cei verificați.

Formula probabilității extragerilor fără revenire (Distribuția Hipergeometrică): Fie $N = N_1 + N_2$.

$$\mathbb{P}(n_1) = \frac{\binom{N_1}{n_1} \cdot \binom{N_2}{n-n_1}}{\binom{N}{n}}$$

Notăm cu $p = \frac{5}{100} = \frac{1}{20}$ probabilitatea ca un student să fie fraudulent.

a) Repartiția variabilei aleatoare X: Variabila aleatoare X (numărul de studenți depistați cu fraudă în $n = 3$ verificări) urmează o distribuție Hipergeometrică:

$$X \sim \text{Hip}(N = 100, M = 5, n = 3)$$

Funcția de probabilitate este $P(X = k) = \frac{\binom{5}{k}\binom{95}{3-k}}{\binom{100}{3}}$, pentru $k \in \{0, 1, 2, 3\}$.

Probabilitățile posibile sunt:

- $p_0 = \mathbb{P}(X = 0) = \frac{\binom{5}{0}\binom{95}{3}}{\binom{100}{3}}$

(Continuarea calculului necesită evaluarea combinațiilor).

—

1.2 Distribuția Binomială Negativă (Pascal)

Distribuția binomială negativă $X \sim \mathrm{NB}(r, p)$ descrie două scenarii legate de un experiment Bernoulli repetat.

Definiția 1: Numărul de Succese înainte de r Eșecuri (sau Succese)

Această definiție este cea pe care ați menționat-o (contorizarea numărului de succese X până la apariția unui număr fixat de eșecuri r).

- Este adecvată atunci când observăm șirul repetărilor până când un număr fixat anterior de **eșecuri** (sau succese) r apar.
- Dacă X este numărul de succese, iar r este numărul de eșecuri fixat (cu probabilitatea succesului p), variabila aleatoare X va avea funcția de probabilitate:

$$P(X = k) = \binom{k+r-1}{k} p^k (1-p)^r, \quad k = 0, 1, 2, \ldots$$

unde $\binom{k+r-1}{k}$ este același lucru cu $\binom{r+k-1}{r-1}$.

X contorizează numărul de succese înainte de a obține r eșecuri.

Definiția 2: Numărul de Încercări necesare pentru al r-lea Succes

Această definiție este, de asemenea, foarte comună și este adesea cunoscută sub numele de **Distribuția Pascal**.

- În acest caz, se dorește contorizarea numărului total de **încercări** necesare pentru a produce al r-lea succes. Notăm acest număr cu X.
- Pentru a obține al r-lea succes exact la a k-a încercare, trebuie să avem $(r-1)$ succese în primele $(k-1)$ încercări, iar a k-a încercare trebuie să fie un succes.
- Funcția de probabilitate (masa de probabilitate):

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, r+2, \ldots$$

Interpretare: Relația de mai sus se citeste: **"Probabilitatea ca la a k-a încercare să obținem al r-lea succes este..."**

2 Exercitii recapitulative variabile aleatoare continue

2.1 Exercițiu Distribuția Normală

Exercițiul 25. Gigel este chemat în instanță pentru a recunoaște paternitatea asupra copilului Lucicăi. Acesta se apără spunând că nu poate fi tatăl copilului întrucât a părăsit țara cu **290 de zile** înainte de nașterea copilului și a revenit în țară cu **240 de zile** înainte de nașterea copilului.

Un expert depune mărturie în cadrul procesului și afirmă că durata sarcinii unei femei este o variabilă aleatoare a cărei repartiție poate fi aproximată cu repartiția normală de **medie** $\mu = 270$ și **dispersie** $\sigma^2 = 100$.

Ce va decide instanța?

Demonstrație. Fie X variabila aleatoare ce reprezintă durata sarcinii. $X \sim \mathcal{N}(\mu = 270, \sigma^2 = 100)$. Deviația standard este $\sigma = \sqrt{100} = 10$.

Gigel este tatăl copilului doar dacă durata sarcinii X a căzut în intervalul $[240, 290]$ de zile. Probabilitatea ca Gigel **să nu fie tatăl** (apărarea sa fiind validă) este probabilitatea ca X să cadă în afara acestui interval: $P((X < 240) \cup (X > 290))$.

$$P((X < 240) \cup (X > 290)) = P(X < 240) + P(X > 290)$$

Prin standardizare ($Z = \frac{X-\mu}{\sigma} = \frac{X-270}{10}$):

$$P(X < 240) = P\left(\frac{X - 270}{10} < \frac{240 - 270}{10}\right) = P(Z < -3)$$

$$P(X > 290) = P\left(\frac{X - 270}{10} > \frac{290 - 270}{10}\right) = P(Z > 2)$$

Calculul final:

$$P((X < 240) \cup (X > 290)) = P(Z < -3) + P(Z > 2)$$

$$= P(Z < -3) + (1 - P(Z \leq 2))$$
$$= \Phi(-3) + (1 - \Phi(2))$$

Folosind valorile tabelelor standard pentru $\Phi(Z)$:

$$\Phi(-3) \approx 0.00135$$

$$\Phi(2) \approx 0.97725$$

$$P_{\text{exterior}} \approx 0.00135 + (1 - 0.97725) = 0.00135 + 0.02275 = \mathbf{0.0241}$$

Probabilitatea ca Gigel să fie tatăl (sarcina să fi durat între 240 și 290 de zile) este $P_{\text{interior}} = 1 - 0.0241 = 0.9759$. Deoarece probabilitatea ca Gigel să nu fie tatăl este extrem de mică (aproximativ 2.41%), instanța va decide că **apărarea lui Gigel nu este plauzibilă**.

—

2.2 Exercițiu Cuplu de Variabile Aleatoare Continue

Se consideră cuplul de variabile aleatoare (X, Y) a cărui repartiție este definită în **triunghiul** care are ca vârfuri originea $\mathbf{O(0,0)}$, punctul $\mathbf{A(0,1)}$ și punctul $\mathbf{B(1,1)}$, prin densitatea de repartiție:

$$f_{X,Y}(x,y) = \frac{a}{\sqrt{xy}}$$

Domeniul de integrare $\mathcal{D}$ (triunghiul OAB) este definit de inegalitățile:

$$0 \le x \le y, \quad 0 \le y \le 1$$

a) Să se determine $a \in \mathbb{R}$.

b) Să se determine densitatea marginală a lui X și a lui Y și să se calculeze $E[X]$, $E[Y]$, $Var[X]$ și $Var[Y]$.

Rezolvare.

a) Determinarea constantei a

Pentru ca $f_{X,Y}(x,y)$ să fie o funcție de densitate de probabilitate, integrala ei dublă pe domeniul $\mathcal{D}$ trebuie să fie egală cu 1:

$$\iint_{\mathcal{D}} f_{X,Y}(x,y)\,dx\,dy = 1$$

$$a \iint_{\mathcal{D}} \frac{1}{\sqrt{x}\sqrt{y}}\,dx\,dy = 1$$

Integrăm în raport cu x de la 0 la y, apoi în raport cu y de la 0 la 1:

$$a \int_0^1 \left(\frac{1}{\sqrt{y}} \int_0^y x^{-1/2}\,dx \right) dy = 1$$

Calculăm integrala interioară:

$$\int_0^y x^{-1/2}\,dx = \left[\frac{x^{1/2}}{1/2} \right]_0^y = [2\sqrt{x}]_0^y = 2\sqrt{y} - 0 = 2\sqrt{y}$$

Substituim rezultatul în integrala exterioară:

$$a \int_0^1 \frac{1}{\sqrt{y}} \cdot (2\sqrt{y})\,dy = 1$$

$$a \int_0^1 2\,dy = 1$$

Calculăm integrala finală:

$$a[2y]_0^1 = 1$$

$$a(2 \cdot 1 - 2 \cdot 0) = 1$$

$$2a = 1 \quad \Rightarrow \quad \mathbf{a} = \frac{\mathbf{1}}{\mathbf{2}}$$

b) Densitățile marginale

Densitatea marginală $f_Y(y)$: Integrăm $f_{X,Y}(x,y)$ pe domeniul lui x. Pentru un y fixat, $x \in [0, y]$:

$$f_Y(y) = \int_0^y f_{X,Y}(x,y)\,dx = \frac{1}{2}\int_0^y \frac{1}{\sqrt{x}\sqrt{y}}\,dx$$

$$f_Y(y) = \frac{1}{2\sqrt{y}}\int_0^y x^{-1/2}\,dx$$

Din punctul a), știm că $\int_0^y x^{-1/2}\,dx = 2\sqrt{y}$.

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot (2\sqrt{y}) = \mathbf{1}, \quad \text{pentru } 0 \le y \le 1$$

Deci, $Y \sim \text{Uniform}[0,1]$.

Densitatea marginală $f_X(x)$: Integrăm $f_{X,Y}(x,y)$ pe domeniul lui y. Pentru un x fixat, $y \in [x, 1]$:

$$f_X(x) = \int_x^1 f_{X,Y}(x,y)\,dy = \frac{1}{2}\int_x^1 \frac{1}{\sqrt{x}\sqrt{y}}\,dy$$

$$f_X(x) = \frac{1}{2\sqrt{x}}\int_x^1 y^{-1/2}\,dy$$

Calculăm integrala:

$$\int_x^1 y^{-1/2}\,dy = [2\sqrt{y}]_x^1 = 2\sqrt{1} - 2\sqrt{x} = 2(1-\sqrt{x})$$

Substituim:

$$f_X(x) = \frac{1}{2\sqrt{x}} \cdot 2(1-\sqrt{x}) = \frac{1-\sqrt{x}}{\sqrt{x}} = \frac{\mathbf{1}}{\sqrt{\mathbf{x}}} - \mathbf{1}, \quad \text{pentru } 0 \le x \le 1$$

c) Calculul mediilor și dispersiilor

Pentru $Y \sim$ Uniform$[0,1]$:

$$\mathbf{E[Y]} = \int_0^1 y \cdot f_Y(y)\,dy = \int_0^1 y \cdot 1\,dy = \left[\frac{y^2}{2}\right]_0^1 = \frac{\mathbf{1}}{\mathbf{2}}$$

$$\mathbf{Var[Y]} = \frac{(b-a)^2}{12} = \frac{(1-0)^2}{12} = \frac{\mathbf{1}}{\mathbf{12}}$$

(Alternativ: $E[Y^2] = \int_0^1 y^2\,dy = \frac{1}{3}$. $Var[Y] = E[Y^2] - (E[Y])^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$).

Pentru X (folosind $f_X(x) = x^{-1/2} - 1$):

$$\mathbf{E[X]} = \int_0^1 x \cdot f_X(x)\,dx = \int_0^1 x\left(x^{-1/2} - 1\right)\,dx = \int_0^1 (x^{1/2} - x)\,dx$$

$$E[X] = \left[\frac{x^{3/2}}{3/2} - \frac{x^2}{2}\right]_0^1 = \left[\frac{2}{3}x^{3/2} - \frac{1}{2}x^2\right]_0^1 = \frac{2}{3} - \frac{1}{2} = \mathbf{\frac{1}{6}}$$

Pentru **Var**[**X**], trebuie calculat $E[X^2]$:

$$E[X^2] = \int_0^1 x^2 \cdot f_X(x)\,dx = \int_0^1 x^2\left(x^{-1/2} - 1\right)\,dx = \int_0^1 (x^{3/2} - x^2)\,dx$$

$$E[X^2] = \left[\frac{x^{5/2}}{5/2} - \frac{x^3}{3}\right]_0^1 = \left[\frac{2}{5}x^{5/2} - \frac{1}{3}x^3\right]_0^1 = \frac{2}{5} - \frac{1}{3} = \frac{6-5}{15} = \frac{1}{15}$$

$$\mathbf{Var}[\mathbf{X}] = E[X^2] - (E[X])^2 = \frac{1}{15} - \left(\frac{1}{6}\right)^2 = \frac{1}{15} - \frac{1}{36}$$

$$Var[X] = \frac{12-5}{180} = \mathbf{\frac{7}{180}}$$

—

3 Inequalities

3.1 Inegalitatea lui Jensen

Fie X o variabilă aleatoare (v.a.) și g o funcție.

- Dacă g este o funcție **convexă**, atunci:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

- Dacă g este o funcție **concavă**, atunci:

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

Observație (legătura cu Varianta): Un caz particular util este alegerea funcției convexe $g(x) = x^2$. Aplicând inegalitatea lui Jensen:

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$$

Aceasta este echivalentă cu afirmația că **Varianța este non-negativă**:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$$

3.2 Inegalitatea Cauchy-Schwarz (pentru variabile aleatoare)

Pentru orice două variabile aleatoare X și Y cu $\mathrm{Var}(X) < \infty$ și $\mathrm{Var}(Y) < \infty$ (adică, momentele de ordinul doi sunt finite), avem:

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Versiuni Echivalente:

a) **Versiunea Discretă (pentru sume):**

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

Această versiune corespunde variabilelor aleatoare discrete X și Y care iau valori $\{a_1, \ldots, a_n\}$ și $\{b_1, \ldots, b_n\}$ cu probabilități uniforme $P(X = a_i, Y = b_i) = 1/n$.

b) **Versiunea Spații Hilbert (Produs Scalar):**

$$|\langle X, Y\rangle| \leq \|X\| \cdot \|Y\|$$

unde produsul scalar este definit ca $\langle X, Y\rangle = \mathbb{E}[XY]$, iar norma este $\|X\| = \sqrt{\mathbb{E}[X^2]}$.

3.3 Markov's inequality

Let $X > 0$ be a non-negative random variable with finite expected value $\mathbb{E}[X] < \infty$. Then for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Interpretation

The probability that a non-negative random variable exceeds a given level a decreases at least linearly with a, provided that its mean is finite.

Proof (continuous case)

Assume X has density $p(x)$. Then

$$\mathbb{E}[X] = \int_0^\infty x\, p(x)\, dx.$$

Split the integral:

$$\mathbb{E}[X] = \int_0^a x\, p(x)\, dx + \int_a^\infty x\, p(x)\, dx.$$

Since $x \geq a$ for all $x \geq a$,

$$\int_a^\infty x\, p(x)\, dx \geq \int_a^\infty a\, p(x)\, dx = a\, P(X \geq a).$$

Hence,

$$\mathbb{E}[X] \geq a\, P(X \geq a),$$

which implies

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Discrete case

The same argument works when X is a discrete non-negative random variable.

Example

Let's say X the distribution for the number of cars passing boulevard Iuliu Maniu on Monday, between 15:00 and 15:05. It has been observed that on average, the number of cars passing is 60. What can we say about the probability of 100 passing?

Note: The data is not real.

$$P(X > 100) < \frac{E[x]}{100} = 0.6$$

It's a gross aproximation, we can surely do better than this.

3.4 Chebyshev Inequality

Let X be a random variable with finite second moment,

$$\mathbb{E}[X^2] < +\infty.$$

Then for any $a > 0$,

$$P\left(|X - \mathbb{E}[X]| \geq a\right) \quad \leq \frac{\mathrm{Var}(X)}{a^2}.$$

In words: the tail probability of X decays at least quadratically.

Proof

Apply Markov's inequality to the non-negative random variable $(X - \mathbb{E}[X])^2$ and the constant a^2:

$$P\big((X - \mathbb{E}[X])^2 \geq a^2\big) \quad \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} \quad = \frac{\mathrm{Var}(X)}{a^2}.$$

Since

$$(X - \mathbb{E}[X])^2 \geq a^2 \iff |X - \mathbb{E}[X]| \geq a,$$

we obtain Chebyshev's inequality.

Example

Coming back to our car example, we found out that we have a variance $Var[X] = 20^2$. How does our aproximation change?

We estimate

$$P(X < 20 \cup X > 100) = P(|X - 60| > 40).$$

By Chebyshev:

$$P(|X - 60| > 40) \leq \frac{20^2}{40^2} = \frac{1}{4}.$$

$$P(X > 100) < P(|X - 60| > 40) < \frac{1}{4}$$

If $X \sim \mathcal{N}(60, 20^2)$

For a true normal distribution,

$$P(X < 20 \cup X > 100) \approx 0.045 < \frac{1}{4}.$$

Also,

$$P(X > 100) \approx 0.022 < \frac{1}{4} < \frac{3}{5}.$$

(Chebyshev gives the bound 1/4;   Markov would give the weaker bound 3/5.)

4 Empirical Mean

4.1 Definition

We call the empirical mean of the random variables

$$X_1, X_2, \ldots, X_m \quad \text{independent and identically distributed}$$

the quantity

$$S_m := \frac{X_1 + X_2 + \cdots + X_m}{m}.$$

4.2 Properties

Let $X_1, \ldots, X_m$ be **i.i.d. (independent and identically distributed)** random variables with $\mathbb{E}[X_1^2] < +\infty$. We denote

$$\mu := \mathbb{E}[X_1], \qquad \sigma^2 := \mathrm{Var}(X_1).$$

(i) $\mathbb{E}[S_m] \quad = \dfrac{m\mu}{m} \quad = \mu.$

(ii) $\text{Var}(S_m) \quad = \dfrac{m\sigma^2}{m^2} \quad = \dfrac{\sigma^2}{m}.$

Observation

The empirical mean S_m approximates μ with variance σ^2/m, which approaches 0 as $m \to +\infty$.

4.3 Example

We are given a biased coin. Let

$$X \sim \text{Bernoulli}(p), \qquad P(X=0) = 1-p, \text{ "tails"}, \qquad P(X=1) = p, \text{ "heads"}$$

with unknown parameter p.

Question 1: How do we determine p?

We toss the coin m times. Each toss is modeled by an i.i.d. random variable $X_i \sim$ Bernoulli(p), for $i = 1, \ldots, m$.

Define

$$S_m = \frac{X_1 + X_2 + \cdots + X_m}{m} \quad = \frac{\#\text{Heads}}{m}.$$

This is an estimator for p, since

$$\mathbb{E}[X_i] = p, \qquad \text{Var}(X_i) = p(1-p).$$

Question 2: How do we determine the estimation error?

We apply Chebyshev's inequality to S_m:

$$P\left(|S_m - p| \geq \varepsilon\right) \quad \leq \frac{\text{Var}(S_m)}{\varepsilon^2} \quad = \frac{p(1-p)}{m\varepsilon^2} \quad \leq \frac{1}{4m\varepsilon^2},$$

since $p(1-p) \leq \frac{1}{4}$.

Thus the probability of an estimation error larger than ε is

$$P\left(|S_m - p| < \varepsilon\right) \geq 1 - \frac{1}{4m\varepsilon^2}.$$

This gives a **confidence level** for estimating p with **margin of error** ε.

Question 3: How many times must we toss a fair coin in order to obtain an error margin $\varepsilon = 0.1$ with confidence level $\alpha = 0.95$ (i.e. 95%) ?

$$\mathbb{P}\left(|S_m - p| < \varepsilon\right) \quad \geq 1 - \frac{1}{4m\varepsilon^2} \quad \geq \alpha.$$

Thus,

$$\frac{1}{4m\varepsilon^2} \leq 1-\alpha \quad \Longleftrightarrow \quad 4m\varepsilon^2 \geq \frac{1}{1-\alpha}.$$

$$\Longleftrightarrow \quad m \geq \frac{1}{4(1-\alpha)\varepsilon^2} = \frac{1}{4 \cdot 0.05 \cdot (0.1)^2} = \frac{1}{4 \cdot 0.05 \cdot 0.01} = 500.$$

$$\boxed{m \geq 500}$$

5 Low of Large Numbers (LoLn)

Given X1, X2, ... Xm iid random variables, having E[X1] = μ and Var[X1] = $\sigma^2 <$ inf, then $\forall \epsilon > 0$, $P(|S_m - \mu| > \epsilon) -> 0$ when $m ->$ inf

Intuition: With a lot of experiments, the empirical mean converges towards the theoretical mean(expected value).